Python을 활용한

다국어 텍스트의 분석 및 활용

- 전처리(preprocessing) 과정을 중심으로

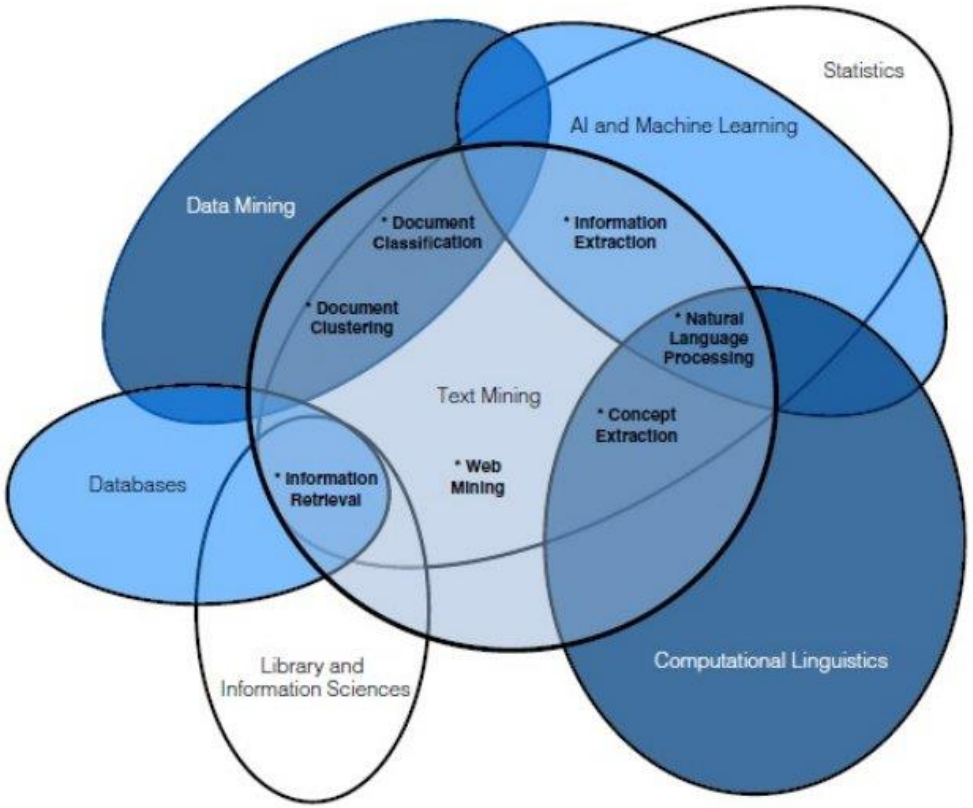
정인정(육군사관학교)

Text Analysis

- 또는 텍스트마이닝(Text Mining):

  텍스트를 대상으로 언어학, 수학, 통계학, 컴퓨터공학 등의 학문적 지식을 이용하여 **특정 목적에 맞게** 유의미한 정보 또는 지식 패턴을 추출하는 분석 및 처리 과정.

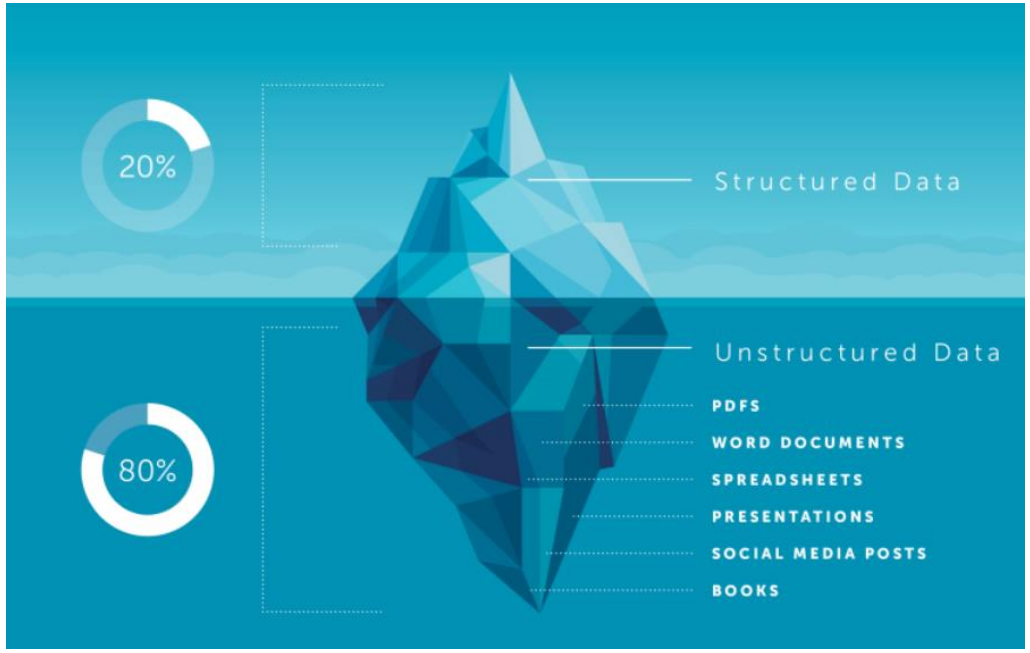
- 많은 텍스트 분석 기법이 자연어처리(Natural Language Processing, NLP)에 기반하고 있음.



Feldman and Sanger(2007), THE TEXT MINING HANDBOOK, Cambridge University Press.

Unstructured Data

- 정형 데이터(Structured Data): (스프레드 시트 형태로) 정해진 규칙에 맞게 들어간 (연산 가능한) 데이터
- 반정형 데이터(Semi-Structured Data): 메타데이터 등을 포함한 데이터. HTML 등.
- 비정형데이터(Unstructured Data): 텍스트 문서(txt), 이미지, 동영상, 블로그, SNS…

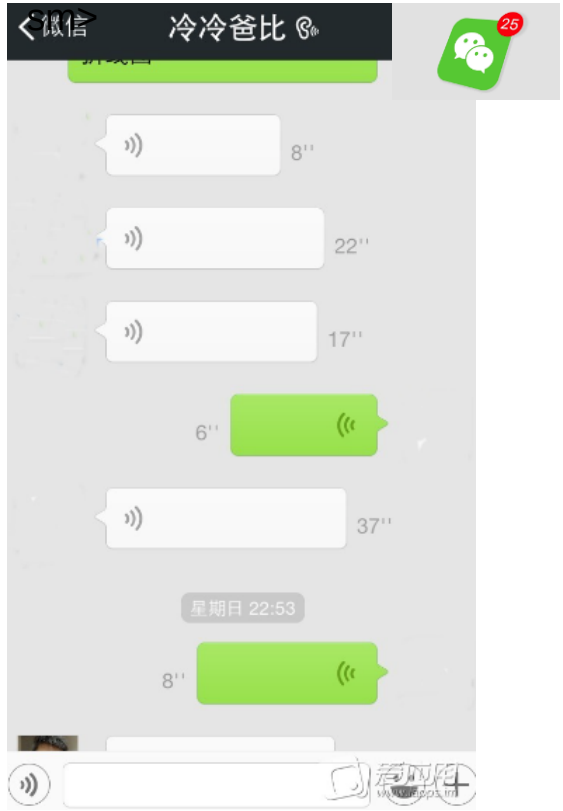


세상에 존재하는 데이터의 80% 이상이 비정형 데이터로 추산
(Chakraborty and Pagolu, 2014)

Text Data의 가치와 활용

- 텍스트 데이터의 폭발적 증가: 매일 늘어나는 데이터의 상당 부분은 텍스트 또는 대화. 웹상에서 사용자들은 주로 텍스트를 활용해 콘텐츠를 생성하고 소통하고 있음.

<중국의 위챗 음성메시지 기능>



<子弹短信(Bullet messenger)의 STT(Speech-To-Text) 기능>

Text Data의 가치와 활용

- 텍스트 데이터 수집의 자동화: Web Crawling, 오픈소스 데이터 등을 통해 데이터의 확보가 용이해짐.

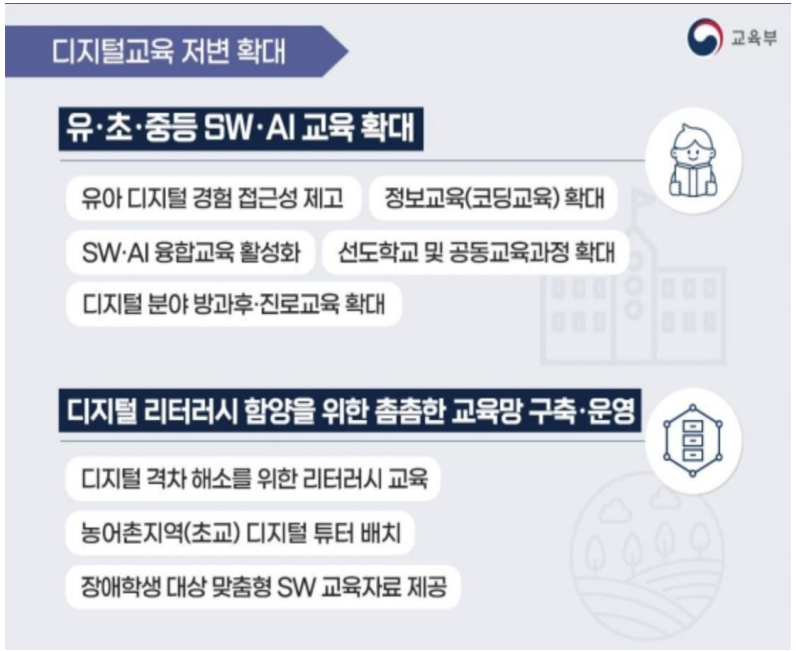
- 최근 사례: 카카오 브레인 초 거대 AI 셋 '코요' 발표 (22.08.31) 약 7억4000만개의 데이터 규모로 직접 개발한 이미지-텍스트를 자동 수집 기능을 활용해 데이터 수집

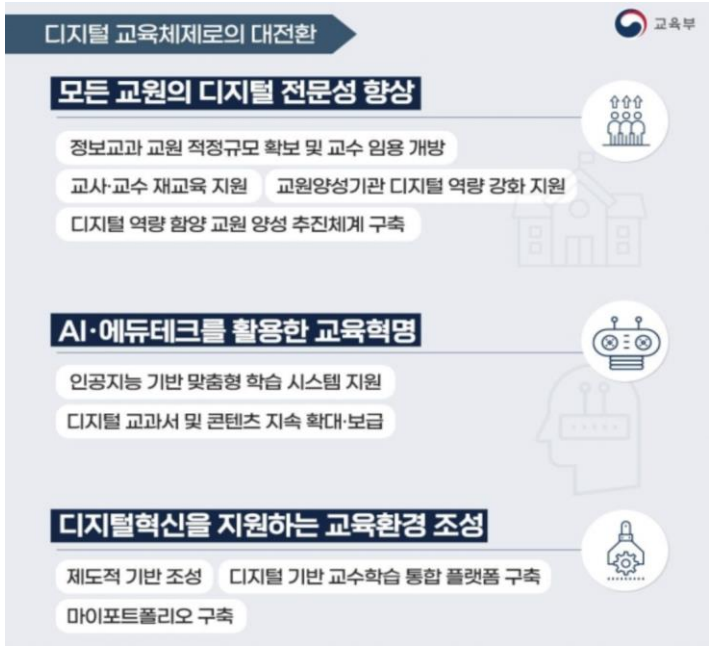
COYO-700M: Image-Text Pair Dataset

Dataset Preview

id	url	text	width	heigh
2559800550877		William Phillips - Fuel State Critical - Outcome in Doubt (B-25	600	300
4896263451343		Fishing Fleet (Monterey), California art by Art Riley. HD giclee art prints for sale at CaliforniaWatercolor.com - original California paintings, & premium giclee prints for sale	600	447

AI와 교육





교육부, 공교육 과정 AI 교육 의무화 계획 발표...

2025년 적용 예정

궁극적 목표: 무엇을 할 것인가?

- 누가 더 좋은, 직관적인 아이디어를 지니고 있는가?

<최신 연구 동향>

김민웅, 황재동 (2021). 한국 다문화 교육에서의 챗봇 활용 가능성 탐구 - 챗봇용 다문화 교육 프로그램 개발 및 효과 분석. 문화교류와 다문화교육, 10(2), 23-42.
이용상, 신동광 (2020). 원격교육 시대의 인공지능 활용 온라인 평가. 학습자중심교과교육연구, 20(14), 389-407.
최원경 (2020). AI 챗봇을 활용한 초등영어 과정중심 말하기 평가: 가능성과 한계. 초등영어교육, 26(1), 131-152.
임희주 (2020). 영어쓰기에서 AWE 프로그램 활용에 대한 대학생의 인식 및 태도 연구. 한국콘텐츠학회논문지, 20(5), 37-45.
박진철 (2021). 인공지능(AI)을 활용한 한국어 듣기 교육 자료 제작 연구 -음성합성기술(TTS) 활용을 중심으로-. 이중언어학, 82, 61-84.
황요한 (2021). 영어교육을 위한 인공지능(AI) 스피커의 활용과 교육적 가치에 관한 연구. 영어영문학연구, 47(1), 225-251.
고권태, 이효영 (2020). 인공지능 챗봇의 중국어 교육 활용 방안 탐색. 중국학, 72, 215-233.
이유희 (2021). 텍스트 마이닝 기법을 통한 일본어능력시험분석과 학습 교육 —JLPT 4급 문자・어휘를 중심으로—. 한국일본문화학회, 91, 283-304.

1차적 목표들: 어떻게 할 것인가?

1. (텍스트)데이터 리터러시(data literacy) 능력의 함양

   데이터를 읽고 이해하며 분석·활용할 수 있는 능력

2. 최신 연구 방법론 및 연구 동향의 이해

   ‘어학/교육학’과 ‘NLP’의 융합적 접점 탐구

   관련 분석 도구 및 연구 논문, 연구 동향 공유

3. 기초적인 파이썬(Python) 능력

   百见不如一打， 百打不如一做

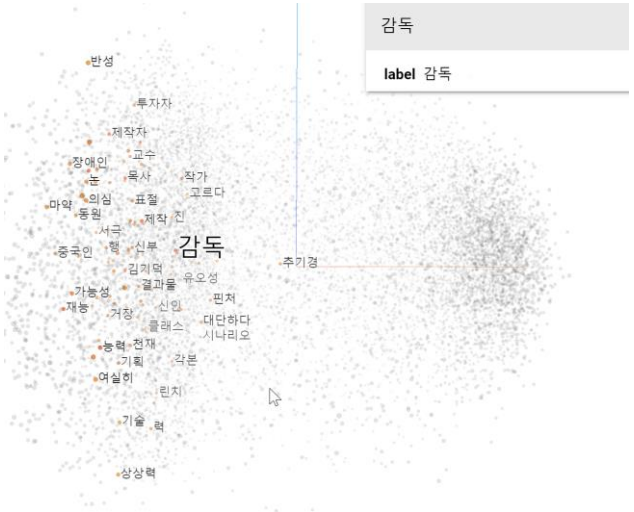
   왜 파이썬인가?

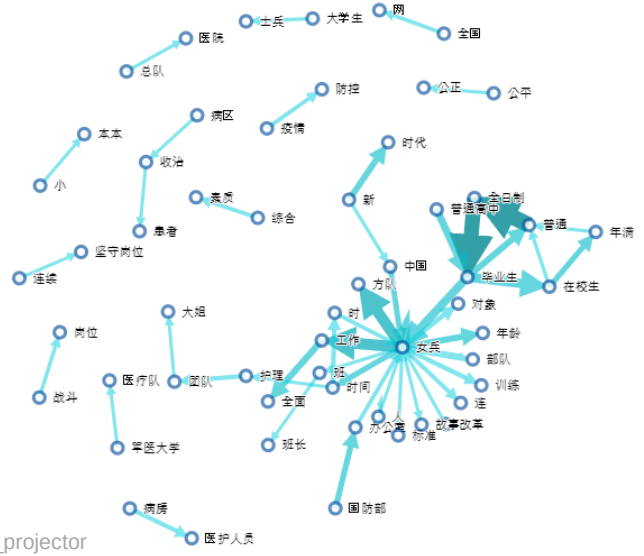
텍스트 분석 기술: 무엇을 할 수 있는가?

빈도 분석

단어간 관계성

단어의 동시 출현 정보





https://soohee410.github.io/embedding_projector

유의미한 단어 추출에서 시작: 텍스트의 전처리(Preprocessing)

텍스트의 전처리

- 텍스트 분석을 위해 공통적으로 수행되는 과정

- 텍스트 또는 문자는 프로그래밍 언어에서 문자열(str)로 표현된다. 컴퓨터에게 텍스트를 이해시키기 위해서는 텍스트를 단어나 형태소 등의 단위로 나눈 후, 리스트(list) 형태로 변환해 주어야 한다.

- 이때 쓸모없다고 생각하는 특수문자나 단어는 제거한다.

```
[ ]  seg_sentence  = jieba.cut("我每天早上喝咖啡") #默认是精确模式
     print(",  ".join(seg_sentence))
```

```
我,  每天,  早上,  喝咖啡
```

텍스트 전처리의 단계

1. 정제(Cleansing): 노이즈 제거, 토큰화 전후로 실행(불용어 제거 등)

2. 토큰화(Tokenization): 텍스트를 단위(Token, 토큰)으로 나누는 작업

3. 정규화(Normalization): 표현 방법이 다른 단어들을 통합시키는 작업

4. 품사 태깅(Tagging): 토큰화한 단어에 대해 품사 부착

1. Cleansing

- html tag, 숫자, 특수문자 등 필요하지 않은 언어 제거

http://news.bookdb.co.kr/bdb/Column.do?_method=ColumnDetail&sc.webzNo=36054&Nnews,"[<p class=""select"">북&피플 , <p class=""select"">칼럼 , ‘사막화의 해답이 된 비눗방울’…과학자의 유희와 사명이 만났을 때 , <p align=""center"" style=""TEXT-ALIGN: center""><span style=""FONT-FAMILY: arial"">

- 문장부호(punctuation) 제거

- 불용어(Stopwords) 제거

2. Tokenization

- 문장 분리(Sentence Tokenization/Sentence Segmentation) 선행

  - 문장부호로 판단

  - 종결 어미 등으로 판단

- 단어 토큰화

  - 공백 기준(영어)

  - 형태소 단위(한국어)

<Konlpy: "아버지가방에들어가신다">

Hannanum	Kkma	Komoran	Mecab	Twitter
아버지가방에 들어가 / N	아버지 / NNG	아버지가방에 들어가신다 / NNP	아버지 / NNG	아버지 / Noun
이 / J	가방 / NNG		가 / JKS	가방 / Noun
시ㄴ다 / E	에 / JKM		방 / NNG	에 / Josa
	들어가 / VV		에 / JKB	들어가신 / Verb
	시 / EPH		들어가 / VV	다 / Eomi
	ㄴ다 / EFN		신다 / EP+EC	

https://konlpy-ko.readthedocs.io/ko/v0.6.0/morph/#pos-tagging-with-konlpy

3. Normalization

- 정규 표현식의 규칙에 따라, 같은 의미의 단어지만 표기가 다른 언어들을 하나의 언어로 정규화한다. 예: USA == US

- 대소문자 통합

- 표제어 추출과 어간 추출 : 단어 원형 찾기

  - 표제어 추출: babies → baby , had → have

  - 어간 추출: fishing, fisher → fish

4. pos-tagging

- 언어별로 다양한 품사 태그표 존재

<중국어 형태소 분석기 jieba의 품사 태그표>

标签	含义	标签	含义	标签	含义	标签	含义
n	普通名词	f	方位名词	s	处所名词	t	时间
nr	人名	ns	地名	nt	机构名	nw	作品名
nz	其他专名	v	普通动词	vd	动副词	vn	名动词
a	形容词	ad	副形词	an	名形词	d	副词
m	数量词	q	量词	r	代词	p	介词
c	连词	u	助词	xc	其他虚词	w	标点符号
PER	人名	LOC	地名	ORG	机构名	TIME	时间

★ 한국어 태그표 비교: https://docs.google.com/spreadsheets/d/1OGAjUvalBuX-oZvZ_-9tEfYD2gQe7hTGsgUpiiBSXI8/edit#gid=0

텍스트 전처리의 필요성

- 성능의 차이는 ‘데이터’의 양과 품질에 달려있다.

- 비규범적 문장과 표현들, 오탈자, 축약어 등의 문제

<Naver sentiment movie corpus v1.0> https://github.com/e9t/nsmc/

id (string)	document (string)	label (class label)
9976970	아 더빙.. 진짜 짜증나네요 목소리	0 (negative)
3819312	흠...포스터보고 초딩영화줄....오버연기조차 가볍지 않구나	1 (positive)
10265843	너무재밋었다그래서보는것을추천한다	0 (negative)
9045019	교도소 이야기구먼 ..솔직히 재미는 없다..평점 조정	0 (negative)
6483659	사이몬페그의 익살스런 연기가 돋보였던 영화!스파이더맨에서 늙어보이기만 했던 커스틴 던스트가 너무나도 이뻐보였다	1 (positive)
5403919	막 걸음마 뗀 3세부터 초등학교 1학년생인 8살용영화.ㅋㅋㅋ...별반개도 아까움.	0 (negative)
7797314	원작의 긴장감을 제대로 살려내지못했다.	0 (negative)
9443947	별 반개도 아깝다 욕나온다 이응경 길용우 연기생활이몇년인지..정말 발로 해도 그것보단 낫겠다 납치.감금만반복반복..이드라마는 가족도없다 연기…	0 (negative)
7156791	액션이 없는데도 재미 있는 몇안되는 영화	1 (positive)

텍스트 전처리의 필요성

- 텍스트의 형태와 특징에 따라 다른 분석방법과 접근 필요

  - 동의어, 동음이의어 문제를 처리할 것인가?

    예: 중국 == 차이나

    IS (Islamic state / International Standard)

  - 토큰화의 단위 설정 문제: 어떻게 쪼갤 것인가?

  - 정제/정규화의 문제: 어디까지 할 것인가??

    예: 문장부호 제거 시 아포스트로피 처리(Don’t) / 영어 소문자 통일 시 축약어 처리 등

개별 언어의 특성

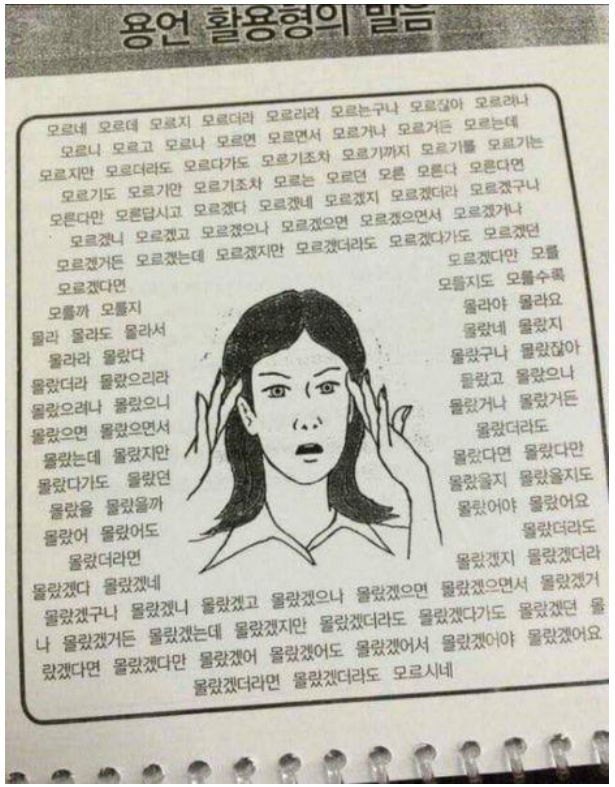
- 텍스트는 전처리가 어렵고, 또 중요한 데이터이며 한국어, 중국어 등은 더욱 그러함.

<한국어> : 교착어의 특성, 띄어쓰기 문제 등

예: 어제 카페 갔었어 → 어제, 카페, 갔었어 / 어제, 카페, 갔어, 어

어제 카페 갔었는데요 → 어제, 카페, 갔었는데요 / 어제, 카페, 갔었, 는데요

(이기창 2019)

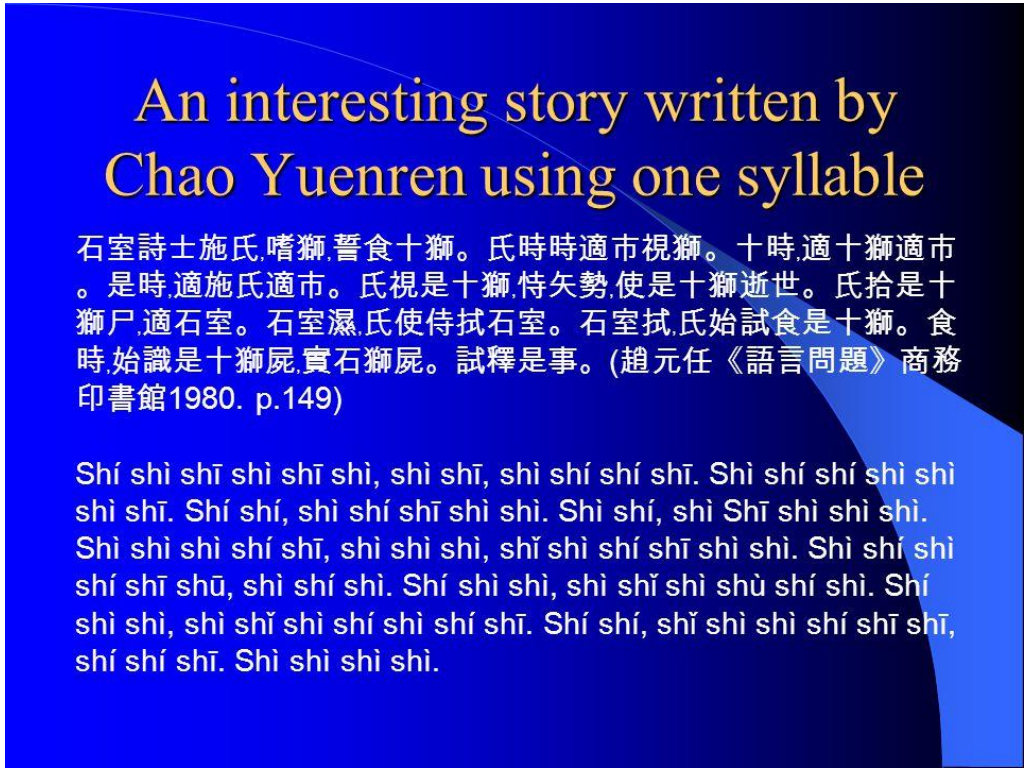


개별 언어의 특성

<중국어>

띄어 쓰기 없음

번체/간체의 문제



텍스트 전처리 도구들

- 텍스트 에디터:

1. notepad++: 윈도우 기반 문서코드/소스코드 편집기. 정규표현식으로 기본적인 전처리 수행 시 유용함. 우측 하단에 인코딩 형식이 제시되어 중국어 데이터 처리 편리.

2. VsCode: MS에서 제공하는 소스코드 편집기

텍스트 전처리 도구들

- 파이썬 오픈소스 라이브러리: Pandas

데이터 처리와 분석을 위한 파이썬 라이브러리. Pandas의 DataFrame은 엑셀의 스프레드시트와 비슷한 테이블 형태를 지닌다.

엑셀로 다루기 불가능한 대용량의 데이터 빠르게 분석 가능.

★ 10 minutes to Pandas https://pandas.pydata.org/pandas-docs/dev/user_guide/10min.html

텍스트 전처리 도구들

- 파이썬 언어 처리 패키지:

1. 영어 NLTK https://www.nltk.org/

NLTK(Natural Language Toolkit)는 교육용으로 개발된 자연어 처리 및 문서 분석용 파이썬 패키지로 말뭉치, 토큰 생성, 형태소 분석, 품사 태깅 등에 사용. xml로 제공, 파이썬에서 nltk.download()로 사용.

2. 한국어 KoNLPy https://konlpy.org/ko/latest/index.html

한국어 정보처리를 위한 파이썬 패키지

텍스트 전처리 도구들

- 언어 처리 패키지:

3. 중국어 Jieba https://github.com/fxsjy/jieba

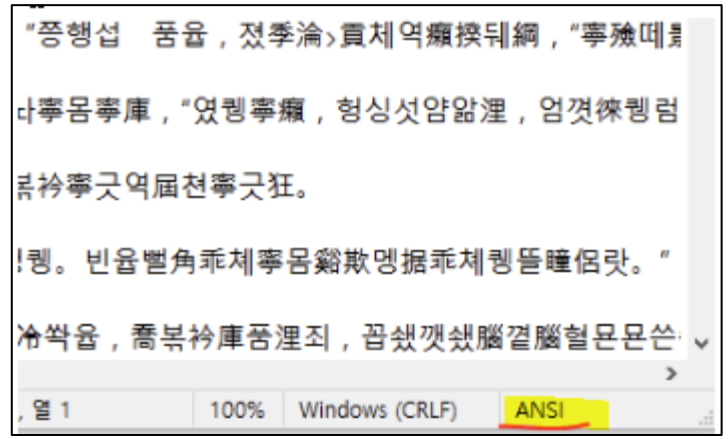
“结巴”中文分词：做最好的 Python 中文分词组件

"Jieba" (Chinese for "to stutter") Chinese text segmentation:

built to be the best Python Chinese word segmentation module.

```
# encoding=utf-8
import jieba
```

전처리 시 주의할 문제들



- 한글/한자의 인코딩 문제 : UTF-8로 인코딩

★ byte와 bit

컴퓨터의 기본 저장 단위는 byte 이며 1byte 는 8bit

- ANSI: 1byte 에는 2의 8승인 256개의 값 저장 가능, 영문자는 따라서 1byte의 ANSI 사용
- UNICODE: 한글, 한자 등은 용량을 크게 확장한 2byte 기반의 UNICODE 사용(2의 16승인 65536개의 고유값 표현 가능)
- UTF-8: UNICODE 포인트를 8bit 숫자의 집합으로 나타내는 것

중국어 텍스트가 gb2312나 gbk로 되어 있으면 그대로 사용할 경우 깨져서 출력되거나 처리 자체가 불가능 하므로 utf-8로의 변환 필요.

전처리 시 주의할 문제들

- 파이썬 자료형 에러(type error) 문제

[] 리스트형

{} 딕셔너리형

str(문자열)과 int(정수) 등.

예제(출처: 웨스트 맥키니, 파이썬 라이브러리를 활요한 데이터 분석, 한빛미디어, 2013)

```
import pandas as pd

data = {'Name': ["John", "Anna", "peter", "Linda"],
        'Location':["New York", "Paris", "Berlin", "London"],
        'Age': [24, 13, 53, 33]
        }

data_pandas = pd.DataFrame(data)

data_pandas
```

	Name	Location	Age
0	John	New York	24
1	Anna	Paris	13
2	peter	Berlin	53
3	Linda	London	33